# MoMask: Generative Masked Modeling of 3D Human Motions

Year: 2023 | Citations: 244 | Authors: Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, Li Cheng

---

## Abstract

We introduce MoMask, a novel masked modeling framework for text-driven 3D human motion generation. In Mo-Mask, a hierarchical quantization scheme is employed to represent human motion as multi-layer discrete motion tokens with high-fidelity details. Starting at the base layer, with a sequence of motion tokens obtained by vector quan-tization, the residual tokens of increasing orders are de-rived and stored at the subsequent layers of the hierar-chy. This is consequently followed by two distinct bidirectional transformers. For the base-layer motion tokens, a Masked Transformer is designated to predict randomly masked motion tokens conditioned on text input at training stage. During generation (i. e. inference) stage, starting from an empty sequence, our Masked Transformer iteratively fills up the missing tokens; Subsequently, a Residual Transformer learns to progressively predict the next-layer tokens based on the results from current layer. Extensive experiments demonstrate that MoMask outperforms the state-of-art methods on the text-to-motion generation task, with an FID of 0.045 (vs e.g. 0.141 of T2M-GPT) on the HumanML3D dataset, and 0.228 (vs 0.514) on KIT-ML, respectively. MoMask can also be seamlessly applied in related tasks without further model fine-tuning, such as text-guided temporal inpainting.